DA5 - INTERPRETING DATA

HSC - General Maths

"Birthdays are good for you.
Statistics show that the people who have the most live the longest"

Larry Lorenzoni

Excellent health statistics - smokers are less likely to die of age related illnesses.'

Name: _____

HSC CAPACITY MATRIX – GENERAL MATHEMATICS

TOPIC: Data Analysis 5 – Interpreting sets of data 2 weeks

CONTENT	CAPACITY BREAKDOWN!	DONE IT!!!!	GOT IT!!!!!	ON MY WAY!	WORKING ON IT!	HELP!!!!
1. Identifies measures of location, ie mean, median 2. Identifies measure of spread ie, range,interquartile range and standard deviation 3. Investigating outliers in small data sets and their effects on the mean, median and mode	Ex 4A Q4, 6, 8- 21					
4. Describing the general shape of a graph or display eg smoothness, symmetry, modes, skew, shape	Ex 4B					
5. Displaying data in back to back stem and leaf plots 6. Displaying data in 2 box and whisker plots on the same scale	Box and Whisker W/S Ex 4C Q1-8					
7. Displaying 2 sets of data on a radar chart 8. Comparing summary statistics from two sets of data	Ex 4C Q9-10 Ex 4D Q 4					
9. Preparing an area chart to illustrate and compare different sets of data over time 10. Comparing summary statistics from two sets of data	Ex 4C Q11 Ex 4D Q5 W/S Further comparative displays					

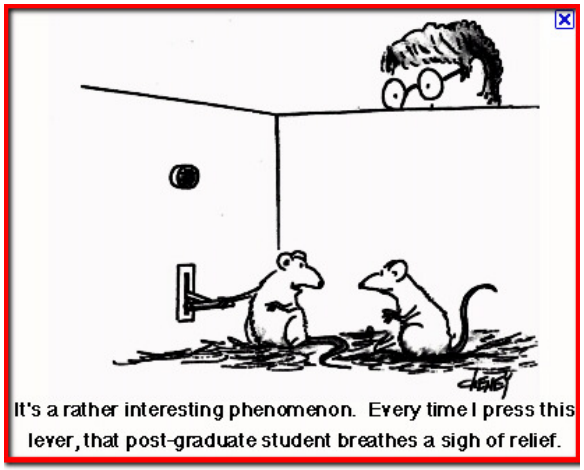
11. Using multiple displays to describe and interpret the relationships between data sets	Ex 4C Q 12 Ex 4D Q1-3, 9-12, 14, 15					
12. Interpreting data presented in two-way table form 13. Comparing summary statistics from two sets of data	Ex 4D Q6-8, 13, 16, 17					

Your say!

What was the most important thing you learned?_____

What was something new you learnt? _____

What part(s) of this topic will you need to work on? _____



rat_cartoon.jpg +1
www2.smumn.edu
Similar - More sizes

THE FACTS SO FAR...

Let's clarify Data to date – can you do AND do you understand the following:

1. For the following data set, calculate the mean, mode and median.

13, 15, 17, 21, 23, 25

a) mean = _____

b) mode = _____

c) median = _____

d) range = _____

e) standard deviation = _____

d) If each score was doubled, how would that impact on:

(i) the mean _____

(ii) the mode _____

(iii) median _____

(iv) range _____

(vii) standard deviation _____

e) If 5 was added to each score, how would that impact on:

(i) the mean _____

(ii) the mode _____

(iii) median _____

(iv) range _____

(vii) standard deviation _____

f) Comment on each question:

(i) Adding or subtracting a value to each element of the data set will affect the following in what way?

(i) the mean _____

(ii) the mode _____

(iii) median _____

(iv) range _____

(vii) standard deviation _____

g) Comment on each question:

(i) Multiplying a value to each element of the data set will affect the following in what way?

(i) the mean _____

(ii) the mode _____

(iii) median _____

(iv) range _____

(vii) standard deviation _____

2. **Finding a missing score of a data set, given the mean:**

   eg Gav played five games of Basketball, averaging 40 points. If the first four games he scored 35, 42, 63, 28, and 32 respectively, what was his final game score?

3. For the following frequency distributions, calculate the mean , mode, median and standard deviation:

x	f	cf
23	4	
34	14	
54	23	
56	21	
78	18	

(i) the mean _____

(ii) the mode _____

(iii) median _____

(iv) range _____

(vii) standard deviation _____

(vii) Interquartile Range _____

(ix) the lowest value of the 9th decile _____

x	f	fx
24	21	
34	24	
35	34	
46	28	
50	16	

(i) the mean _____

(ii) the mode _____

(iii) median _____

(iv) range _____

(vii) standard deviation _____

(vii) Interquartile Range _____

(ix) the lowest value of the 9th decile _____

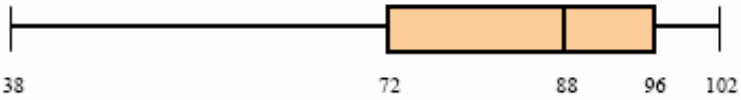
4. When do you use σ_{n-1}?

5. Suppose that one family kept track of how many DVDs they rented each month for a two year period. The numbers for each month are shown in the tble below. Make a box and whisker graph from this data.

J	F	M	A	M	J	J	A	S	O	N	D	J	F	M	A	M	J	J	A	S	O	N	D
3	5	2	8	1	5	0	3	6	4	9	15	3	6	4	1	10	3	8	7	2	9	0	11

6. For questions 1 – 6, refer to the box & whisker graph below which shows the test results of a math class.

Test Scores (as %)



_____ 1. What was the high score on the test?

_____ 2. What percent of the class scored above a 72?

_____ 3. What was the median score on the test?

_____ 4. What percent of the class scored between 88 & 96?

5. Do you think that this test was too hard for the students? Explain.

_____

_____

6. Would you expect the mean to be above or below the median? Explain.

_____

_____

http://illuminations.nctm.org/activitydetail.aspx?ID=160 Excellent site for Box and Whisker plots

MEASURES OF LOCATION

REMEMBER...

To identify a score that is typical of a data set, we can use the mean or median
For a small set of scores,

$$\bar{x} = \frac{\Sigma x}{n} = \frac{\text{sum of all scores}}{\text{number of scores}}$$

For a large set of scores,

$$\bar{x} = \frac{\Sigma fx}{\Sigma f} = \frac{\text{sum of the f.x column}}{\text{sum of frequencies}}$$

The median is the middle score when all scores are in ascending order.
The mode is the most popular data score. There may be more than one mode in a data set.

eg In a street there are five houses. The values of the houses are:

$450 000, $475 000, $475 000, $490 000, $505 000

A new house is built and valued at $ 760 000. Describe the effect that this outlier has on the:

a) mean _____

b) median _____

c) mode _____

eg For the following data set, calculate the mean, mode and median.

13, 15, 17, 21, 23, 25

a) mean = _____

b) mode = _____

c) median = _____

d) If each score was doubled, how would that impact on:

(i) the mean _____

(ii) the mode _____

(iii) median _____

e) If 5 was added to each score, how would that impact on:

(i) the mean _____

(ii) the mode _____

(iii) median _____

eg For the two sets of data, calculate the mean and median:

DATA SET A: 55, 57, 57, 58, 60, 60, 62, 63, 63, 65

$\bar{x} =$ _____

Median = _____

DATA SET B: 13, 19, 31, 40, 55, 65, 90, 92, 95, 100

$\bar{x} =$ _____

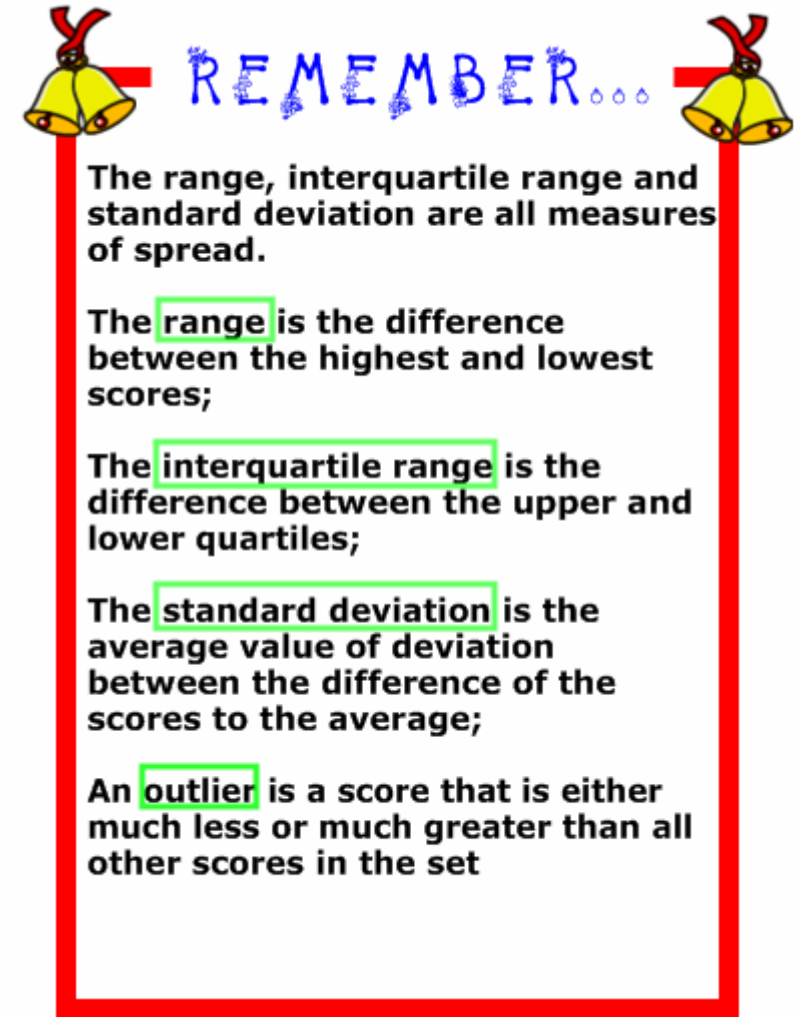
Median = _____

Compare and comment on the measures of location for the two data sets:

_____

chipmunks-christmas-graphic2.jpg
mudtrap.com

MEASURES OF SPREAD



eg Considering the two data sets at the top of the page, calculate the following for each set:

DATA SET A: 55, 57, 57, 58, 60, 60, 62, 63, 63, 65

Range = _____

IQR = _____

σ_n = _____

DATA SET B: 13, 19, 31, 40, 55, 65, 90, 92, 95, 100

Range = _____

IQR = _____

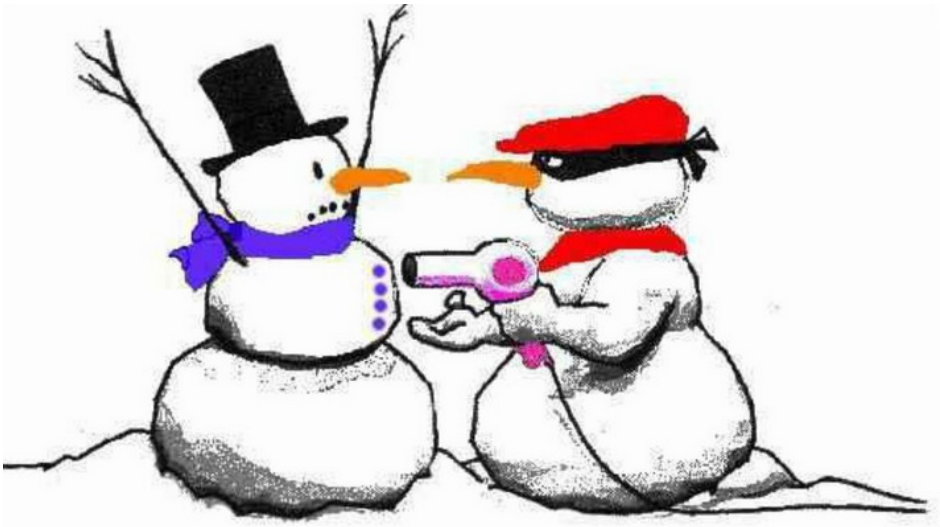
σ_n = _____

Compare and comment on the measures of spread for the two data sets

_____

_____

_____



INVESTIGATING OUTLIERS & THEIR IMPACT

INVESTIGATION: For the following sets of data, calculate the mean, standard deviation, median and interquartile range and construct a box and whisker plot.

SET A: 2, 2, 2, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9, 19

SET B: 2, 2, 2, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 9, 9

SET A		SET B	
Mean		Mean	
Median		Median	
Standard Dev		Standard Dev	
IQR		IQR	

The advantage of using the mean and standard deviation is that they:

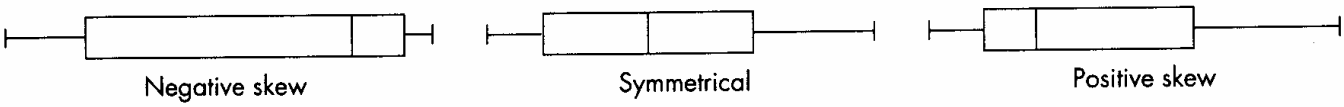
The disadvantage of using the mean and standard deviation is that they:

The advantage of using the median and IQR is that they:

The disadvantage of using the median and IQR is that they:

ANALYSIS OF DATA

One aspect which can be considered is skewness. **Skewness** is, in a sense, the opposite to symmetry. The following box plots illustrate the concept.



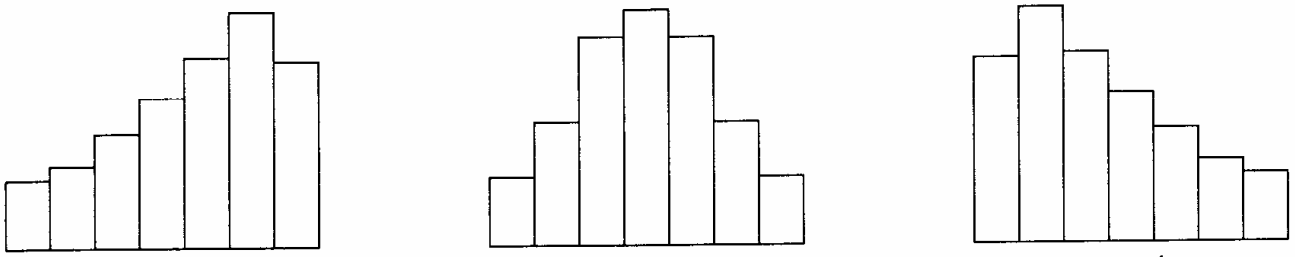
Negative skew Symmetrical Positive skew

In a box plot **negative skew** indicates that the median is closer to the upper quartile than the lower quartile. We sometimes say that the tail to the left is bigger than that to the right.

A **symmetrical** box plot has the median in the middle of the box. Note that this has nothing to do with the size of, or the placement of, the whiskers.

In a box plot **positive skew** indicates that the median is closer to the lower quartile than the upper quartile. We can say the tail to the right is bigger than to the left.

These terms can also be applied to histograms, as shown in the following diagrams.



Negative skew Symmetrical Positive skew

Although not always absolutely accurate, the following *rules of thumb* give us an indication of what is occurring in a skewed data set.

For a positively skewed data set: mode < median < mean

For a negatively skewed data set: mean < median < mode

REMEMBER...

A distribution is SYMMETRICAL when the data are equally distributed around the mean;
When the data is symmetrical, the median and mean will BOTH be the middle score;
When the data are clustered around the mean, the standard deviation is smaller;

When the majority of the scores are at the LOWER END of a distribution, it is POSITIVELY SKEWED;

When the majority of the scores are at the UPPER END of the distribution, it is NEGATIVELY SKEWED.

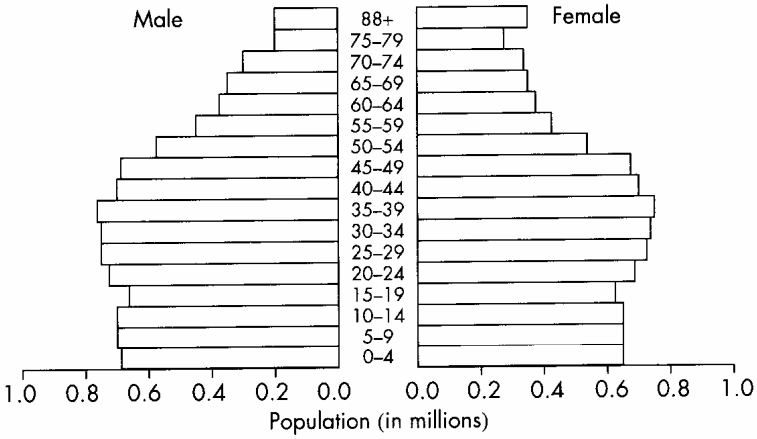
eg The data set below represents the marks obtained on a test. Construct a box and whisker plot and hence describe the following data set.

14, 23, 24, 25, 25, 26, 28, 33, 35, 37, 38, 38, 39, 41, 43, 45, 46, 48, 49, 49

COMPARATIVE DISPLAYS

These can be produced in a number of ways. They may be presented like the age-gender pyramid shown, they may have parts of the bars shaded differently to represent the variables, or they may have different coloured columns side by side. Some types of comparative bar charts can handle more than two categories.

Population of Australia 1997



Back-to-back stem-and-leaf plots

These have a central stem with the leaves moving away from the stem in either direction. A back-to-back stem-and-leaf plot can be used to compare only two categories.

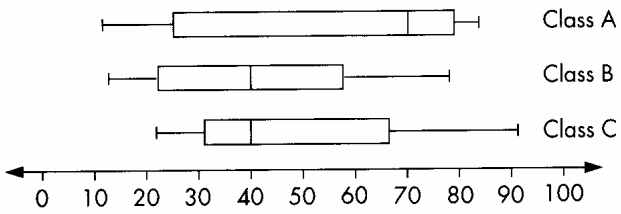
Hockey players' heights (cm)

Team A		Team B
9 8 5	15	1 4 6 7
6 5 4 4 1	16	0 8
9 8 8 6 5	17	2 4 4 6
	18	3 6
	19	0

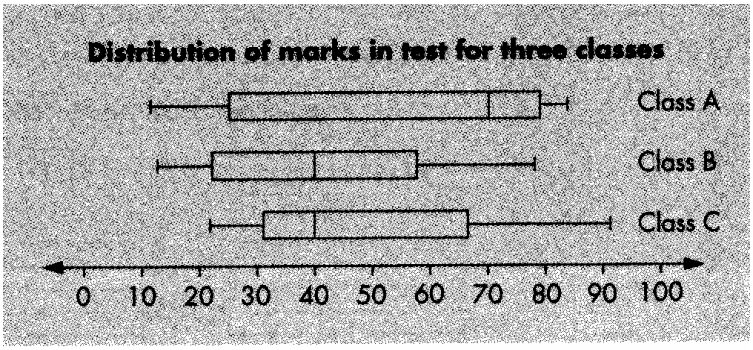
Comparative box plots

These have several related box plots drawn using the same scale. A comparative box plot as shown in Worked Example 9 can be used for any number of categories.

Distribution of marks in test for three classes

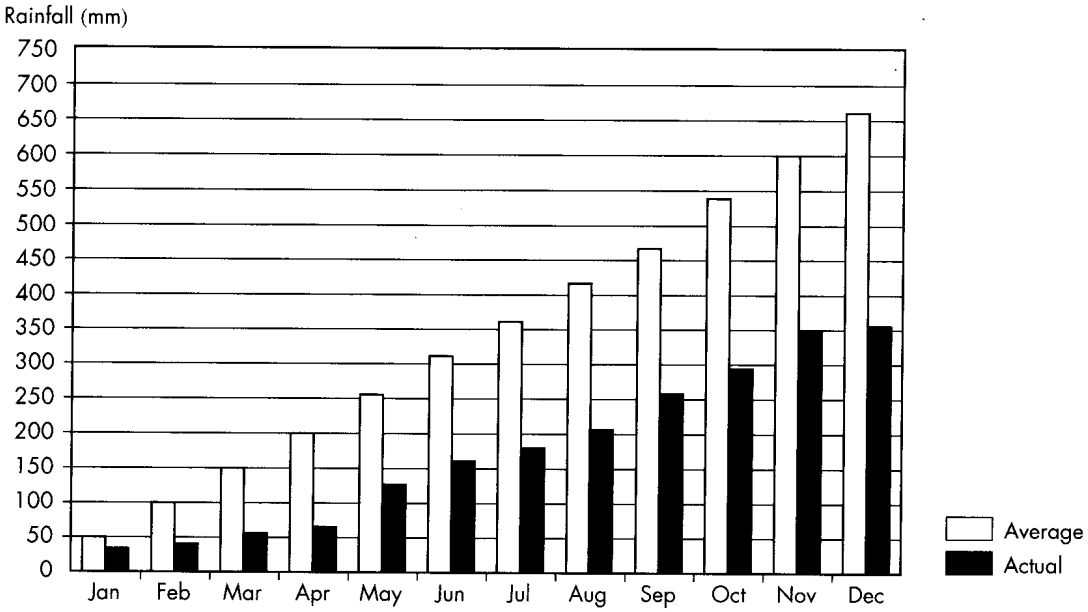


We have described individual data sets in terms of their degree of skewness, their type of skewness, measures of central tendency and measures of spread. We can use these descriptive features when comparing data sets, especially when we wish to refer to specific similarities and differences.



eg Write a comparative description of the data sets, noting as many relevant points as possible,

The weather in Melbourne is notoriously variable. The following bar chart compares the year-to-date rainfall, per month, in Melbourne for 1997 against the average rainfall. Discuss the content of the chart.



The following back-to-back stem-and-leaf plot shows the ages of the winners of the best actor and best actress Academy Awards for the years 1928–98.

BEST ACTOR		BEST ACTRESS
	2_L	1 2 4 4 4 4 4
	2_H	5 6 6 6 6 6 6 6 7 7 7 8 8 9 9 9 9
4 4 3 3 2 2 2 1 1 0	3_L	0 0 0 0 1 1 2 3 3 3 3 4 4 4 4 4 4 4
9 9 8 8 8 8 8 8 7 7 7 5 5 5 5	3_H	5 5 5 6 7 7 8 8 8 9
4 4 3 3 3 3 3 2 2 2 1 1 1 1 0 0 0 0	4_L	0 1 1 1 1 1 1 2
9 9 9 8 8 8 7 7 6 6 5 5	4_H	5 5 8 9 9
4 3 2 2 1 1	5_L	
6 6 6 5	5_H	
2 1 0 0	6_L	0 1 1 2
	6_H	
	7_L	4
6	7_H	
	8_L	0

Comment on the two distributions, including a discussion of measures of central tendency, measures of spread and the degree of skewness. Draw a comparative boxplot to help with the analysis.

eg

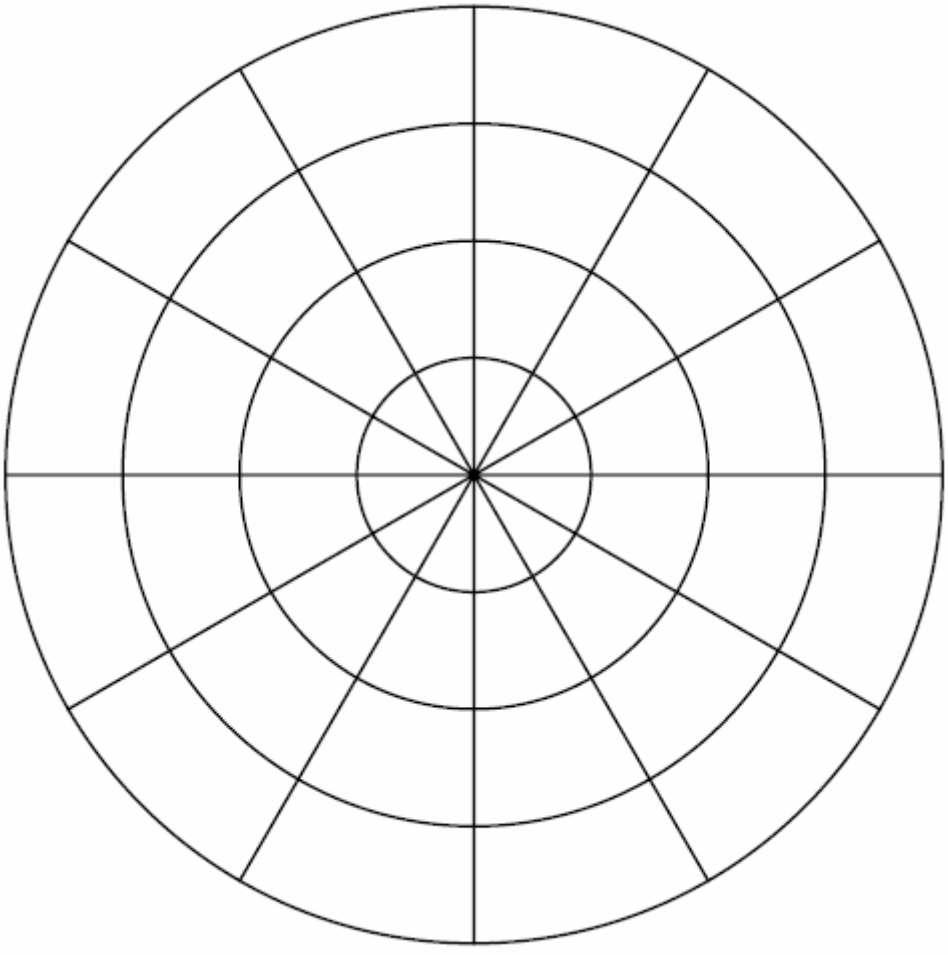
The data shown below display the marks of 15 students in both English and Maths.

English:	**45**	**67**	**81**	**59**	**66**	**61**	**78**	**71**	**74**	**91**	**60**	**49**	**58**	**62**	**70**
Maths:	**85**	**71**	**49**	**66**	**64**	**68**	**75**	**71**	**69**	**60**	**63**	**80**	**87**	**54**	**59**

Display the data in a back-to-back stem-and-leaf plot.

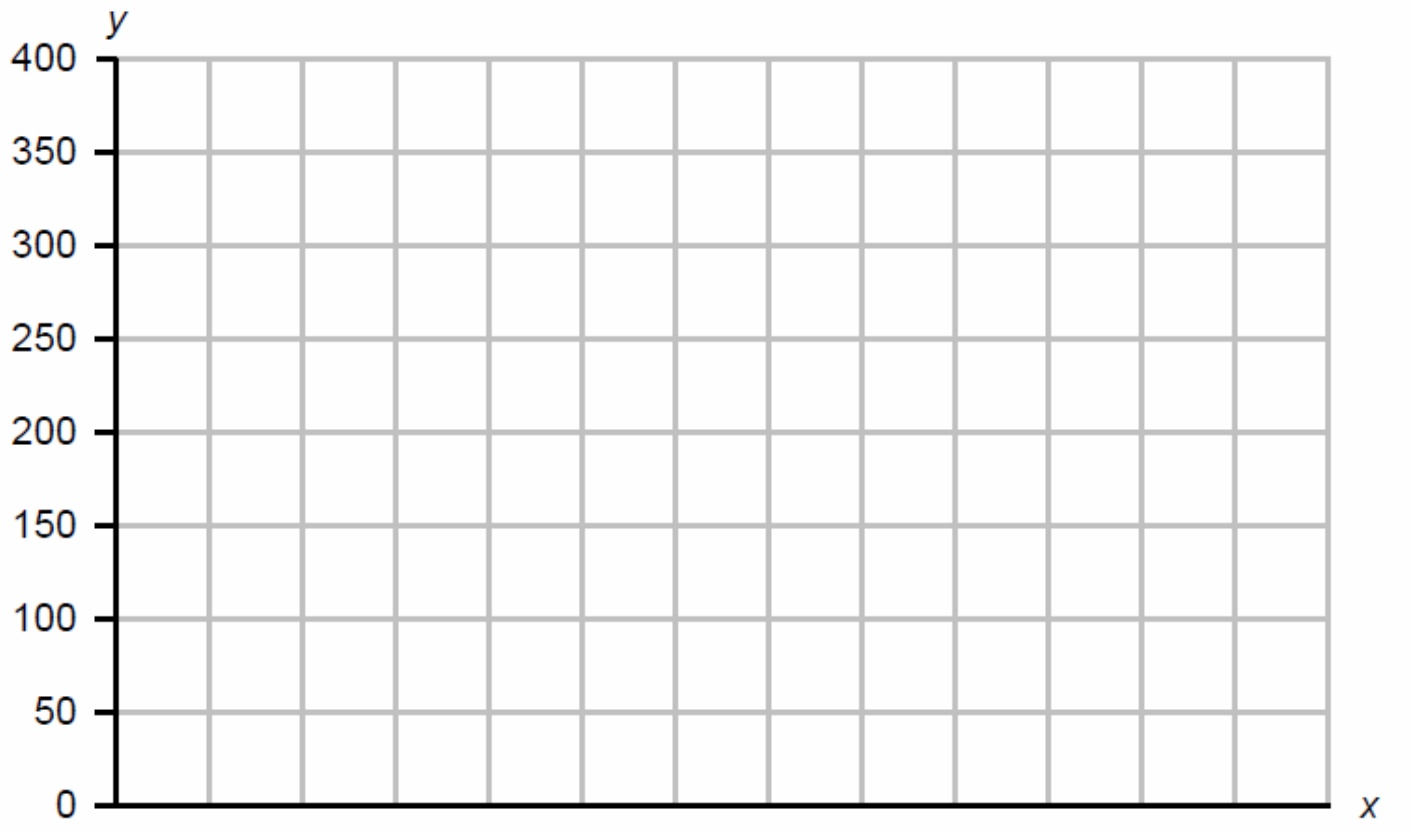
eg **The table below shows the number of admissions to two hospitals, each month, over a one-year period. Display both sets of data on a radar chart.**

Month	Hospital A	Hospital B
January	3	15
February	6	12
March	7	9
April	9	10
May	10	8
June	15	7
July	14	9
August	16	6
September	10	8
October	5	5
November	3	9
December	7	2



eg The table below shows the amount of rainfall, in millimetres, in Sydney, Melbourne and Brisbane each month throughout the year. Display this information in an Area chart.

	January	February	March	April	May	June
Sydney	103	117.1	133.7	126.6	120.4	131.7
Melbourne	49	47.7	51.8	58.4	57.2	50.2
Brisbane	159.6	158.3	140.7	92.5	73.7	67.8
	July	**August**	**September**	**October**	**November**	**December**
Sydney	98.2	79.8	69.9	77.5	83.1	79.6
Melbourne	48.7	50.6	59.4	67.7	60.2	59.9
Brisbane	56.5	45.9	45.7	75.4	97	133.3



RADAR CHARTS AND AREA CHARTS

RADAR CHARTS:

A REMINDER: When drawing a radar chart it is necessary to calculate the number of degrees between the arms of the graph, using:

$$\text{number of degrees} = \frac{360^\circ}{\text{number of categories}}$$

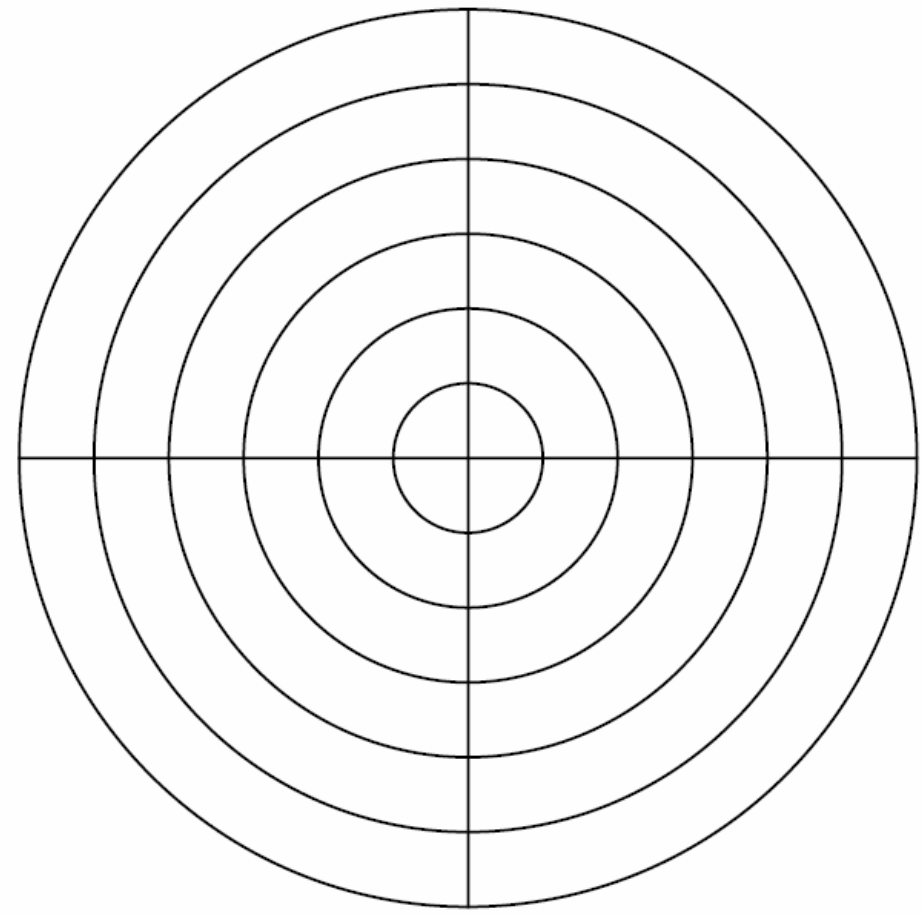
eg Construct a radar chart to compare the quarterly takings for two operators in the CleanUp rubbish removal franchise.

OPERATOR	QUARTER 1	QUARTER 2	QUARTER 3	QUARTER 4
Wright's Rubbish	$3200	$4000	$3300	$5250
Lightly's Litter	$4120	$3800	$3500	$2500

THINK:

1. *Calculate the number of degrees between the arms of the chart using the above formula*

Construct the chart using a suitable scale for the axis



AREA CHARTS:

An area chart is best used to display the magnitude of change over time of various quantities. As the chart also displays the sum of the values plotted it also gives us a feel for the relationship of the individual parts of the whole.

eg The table shows the number of recorded offences related to the possession of drugs in NSW over a four-year period.

OFFENCE	2001	2002	2003	2004
Possession and use of cocaine	137	117	152	273
Possession and use of narcotics	1296	1541	1849	2977
Possession and use of cannabis	9060	9742	10 252	11 159
Possession and use of other drugs	1109	1052	1290	1787

Bureau of Crime Statistics and Research NSW Australia

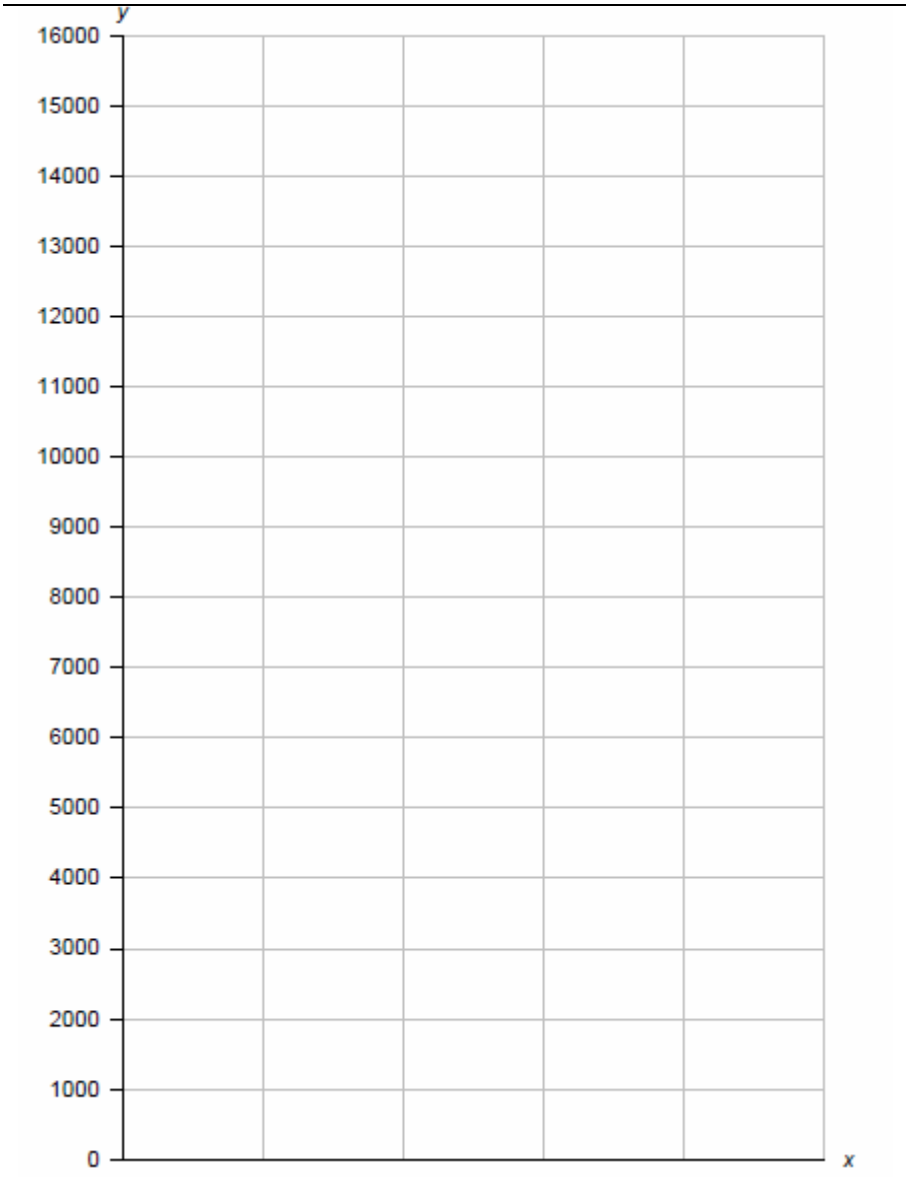
Construct an area chart to present this data and write a brief comment about what it tells you.

THINK:

1. *Find progressive totals for each column. This is because each line the graph takes the line below it as its baseline.*

	2001	2002	2003	2004
Cocaine	137	117	152	273
Cocaine + Narcotics				
Cocaine + Narcotics + Cannabis				
Cocaine + Narcotics + Cannabis + Other				

2. *Construct the graph, plotting each set of values individually. (Note that the number of cocaine offences is so small it is difficult to see them on the graph.*



3. *Comment on the graph*

COMMENT: